Improving interaction prediction exploiting background information

Konstantinos Pliakos KONSTANTINOS.PLIAKOS@KULEUVEN.BE
KU Leuven Kulak, Department of Public Health and Primary Care, Etienne Sabbelaan 53, 8500 Kortrijk, Belgium

Celine Vens CELINE.VENS@KULEUVEN.BE
KU Leuven Kulak, Department of Public Health and Primary Care, Etienne Sabbelaan 53, 8500 Kortrijk, Belgium

Keywords: Interaction prediction, Random Forest, kernel PCA

Abstract

During the last years, a burst of interest has been witnessed in the prediction of interactions that occur in biomedical networks. Despite the research effort made so far, accuracy and efficiency are still open problems. Here, a new prediction scheme is proposed that is based on supervised learning using Random Forest (RF) extended by Kernel Principal Component Analysis (KPCA). The obtained experimental results reaffirmed the potential of the proposed approach.

1. Introduction

Nowadays, the amount of data used in various areas of science increases exponentially. This vast amount of data is followed by more complex representations, further referred to as structured data types. One example of structured data that is often encountered is interaction data. The main point is that instead of one set of objects described by a set of features, interaction data is characterized by two sets of objects, each described by its own set of features. In this framework, the goal is to predict the interactions between both object sets, a process also referred to as network inference. Interaction data is omni-present: in social network analysis, recommender systems, ecology (habitat modeling), bioinformatics (gene expression analysis, drug response analysis, predicting drug-target reactions), technology-enhanced education, etc. Despite the fact that recent technological advances have increased the number of known interactions in networks, such as protein-protein (PPI) or drug-protein (DPI) networks, these interactions cover only a fraction of the complete corresponding networks. In addition to that, the experimental methods for identifying such interactions are both expensive and time-consuming. To this end, there is an indisputable need of developing new computational methods to efficiently predict interactions.

Many works have focused on interaction prediction in complex networks. In (Yu et al., 2012), a systematic approach efficiently integrating the chemical, genomic, and pharmacological information for drug targeting and discovery was developed. It was based on Random Forest (RF) and Support Vector Machine (SVM). In (You et al., 2013), an hierarchical PCA-EELM (principal component analysis-ensemble extreme learning machine) model was proposed to predict PPI interactions based on the information of protein sequences. More recently, a study (Schrynemackers et al., 2015) presenting three tree-based ensemble methods for biological network inference was presented.

Here, motivated by (Schrynemackers et al., 2015), a method is proposed where interactions are predicted using tree-based ensemble methods and KPCA. In particular, KPCA is employed as an effective feature extraction technique, generating a more discriminative feature set from the original one. By reducing the dimensions the existing noise in the dataset is discarded and the power of the new feature set can also be boosted by providing a suitable kernel. Finally, tree-based ensembles are applied on the concatenation of the new feature set and the original one. This way, the performance is enhanced and the interpretability, by means of feature ranking, provided by the tree-based methods, is kept.

Appearing in *Proceedings of Benelearn 2016*.

2. Method

Interaction frameworks are often presented as networks or graphs. More specifically, the interaction sets define the nodes of a graph and the interactions between these nodes are represented in the adjacency matrix of this graph. Let $\mathbf{Y}$ be that adjacency matrix, having entries $y_{ij} = 1$ in case of an existing interaction between node i and j and 0 otherwise. The samples in the two interaction sets defined as the nodes are described by their feature representation (i.e., feature vectors). Let $\mathbf{X}$ be that feature set.

Here, the interaction prediction task is based on a supervised learning framework using multi-output tree-based ensembles trained on $\mathbf{X}$ (Schrynemackers et al., 2015). This way, interactions are predicted at an inductive manner, exploiting the background information of the interaction sets. Denoting by f a prediction function, the interaction prediction is modelled as $f(X) \rightarrow Y$.

In order to improve the efficacy of the method, a new feature set $\mathbf{F}$ is generated from the original one, using KPCA. KPCA is one approach of generalizing linear PCA into non-linear case using the kernel method. The original feature vectors $\mathbf{X}_i$ of $\mathbf{X}$ are initially mapped into a high-dimensional feature space $\mathbf{\Phi}(\mathbf{X}_i)$ and then the linear PCA is calculated in $\mathbf{\Phi}(\mathbf{X}_i)$. This way, the linear PCA in $\mathbf{\Phi}(\mathbf{X})$ corresponds to a non-linear PCA in $\mathbf{X}$. In addition to exploiting the possible existence of a non-linear manifold in $\mathbf{X}$, by significantly reducing the dimensions it is expected that most of the noise present in the original set is removed. This way, the joint feature set is expected to be more informative and the interpretability advantage of the tree-based ensembles (i.e., feature ranking) is also maintained. The problem of computational complexity is also diminished by applying dimensionality reduction on the new feature set. By denoting d as the number of dimensions kept after KPCA the proposed method is finally modelled as $f(X + F_d) \rightarrow Y$.

3. Experimental results

Despite the many tree-based ensemble prediction methods, here the Random Forest (RF) was employed as it is one of the most well-known predictors. The proposed approach, coined as ERF (Extended RF), was evaluated employing the interaction datasets used in (Schrynemackers et al., 2015) and it was compared to a similar version of the multi-output approach proposed in the same work. These datasets represent both homogeneous and heterogeneous interaction networks. Out of simplicity, the heterogeneous datasets were treated as two separate datasets. Furthermore, the Area Under Precision Recall curve (AUPR) and the Area Under ROC curve (AUROC) were used as measures of merit. The 5-fold cross validation approach was followed and the RF training was performed using 50 trees. As it is shown in Table 1, the proposed approach yields generally better results in terms of AUPR and AUROC. One can further increase the results by tuning all the parameters or by choosing a better kernel than the common polynomial used here. The parameter d was set equal to the square root of the number of dimensions of $\mathbf{X}_i$.

Table 1. AUPR and AUROC values for the two methods.

Data	AUPR RF/ERF	AUROC RF/ERF
PPI(protein-protein)	0.17/0.18	0.80/0.81
MN (enzymes)	0.13/0.13	0.76/0.76
ERN (TF-genes)	0.40/0.42	0.85/0.86
ERN2 (genes-TF)	0.09/0.09	0.65/0.66
DPI (drug-protein)	0.13/0.13	0.75/0.75
DPI2 (protein-drug)	0.03/0.04	0.60/0.62
SRN (TF-genes)	0.20/0.20	0.82/0.82
SRN2 (genes-TF)	0.02/0.02	0.52/0.52

4. Conclusions

Here, an efficient interaction prediction approach was proposed. It is based on multi-output prediction using tree-based ensembles and it was further enhanced by including a more informative feature set in the learning procedure in concatenation with the original one. The potential of the proposed method was demonstrated experimentally and improvements could be brought off by further research.

References

Schrynemackers, M., Wehenkel, L., Babu, M. M., & Geurts, P. (2015). Classifying pairs with trees for supervised biological network inference. *Molecular BioSystems*, *11*, 2116–2125.

You, Z.-H., Lei, Y.-K., Zhu, L., Xia, J., & Wang, B. (2013). Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis. *BMC bioinformatics*, *14*, S10.

Yu, H., Chen, J., Xu, X., Li, Y., Zhao, H., Fang, Y., Li, X., Zhou, W., Wang, W., & Wang, Y. (2012). A systematic prediction of multiple drug-target interactions from chemical, genomic, and pharmacological data. *PloS one*, *7*, e37608.